Retro-predicting language change with binomial regression analysis

Freek Van de Velde
(University of Leuven – QLVL)

- "It's hard to make predictions, especially about the future." (Yogi Berra)

- "It's hard to make predictions, especially about the future." (Yogi Berra)

- There is fairly wide-spread pessimism about the predictability of language change:

  "(...) malgré certaines apparences contraires, les événements diachroniques ont toujours un caractère accidentel et particulier." (De Saussure 1955[1916]: 131)

  "Diachronic linguistics is not a predictive science" (Bauer 1994: 25)

  "The common view is that it is not possible to make sensible predictions about future linguistic developments." (Sanchez-Stockhammer 2007)

- "It's hard to make predictions, especially about the future." (Yogi Berra)

- There is fairly wide-spread pessimism about the predictability of language change:

  "(...) malgré certaines apparences contraires, les événements diachroniques ont toujours un caractère accidentel et particulier." (De Saussure 1955[1916]: 131)

  "Diachronic linguistics is not a predictive science" (Bauer 1994: 25)

  "The common view is that it is not possible to make sensible predictions about future linguistic developments." (Sanchez-Stockhammer 2015)

- Distinction between 'innovation' and 'propagation' (Croft 2000)
  - Innovation. Predicting when a change will occur is next to impossible (the actuation problem Weinreich, Herzog & Labov 1968)
  - Propagation. Predicting the trajectory of the change through a community through time can be modelled (Blythe & Croft 2012)

- "It's hard to make predictions, especially about the future." (Yogi Berra)

- There is fairly wide-spread pessimism about the predictability of language change:

  "(...) malgré certaines apparences contraires, les événements diachroniques ont toujours un caractère accidentel et particulier." (De Saussure 1955[1916]: 131)

  "Diachronic linguistics is not a predictive science" (Bauer 1994: 25)

  "The common view is that it is not possible to make sensible predictions about future linguistic developments." (Sanchez-Stockhammer 2015)

- Distinction between 'innovation' and 'propagation' (Croft 2000)
  - Innovation. Predicting when a change will occur is next to impossible (the actuation problem Weinreich, Herzog & Labov 1968)
  - Propagation. Predicting the trajectory of the change through a community through time can be modelled (Blythe & Croft 2012)

  → 'Predicting when someone will drop a glass on the floor is next to impossible, but predicting the glass's downward trajectory, due to gravity can be modelled.

- This talk:
  - Predicting the future, in a falsifiable way, by looking at a former future time point which now lies in the past

- This talk:
  - Predicting the future, in a falsifiable way, by looking at a former future time point which now lies in the past

- This talk:
  - Predicting the future, in a falsifiable way, by looking at a former future time point which now lies in the past

  - Using, and extending, binomial regression
  - Three case studies, in Late Modern Dutch:
    - Hortative
    - Predeterminers
    - Complex prepositions

S-curves

- Robust language changes mostly take the shape of an S-curve (Weinreich et al. 1968: 113; Kroch 1989; Croft 2000; Denison 2003; Pintzuk 2003; Blythe & Croft 2012; Nevalainen 2015)

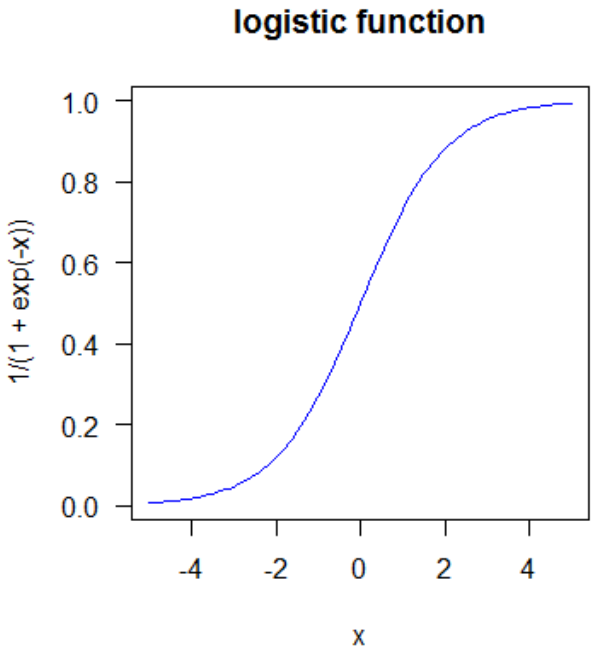
- This shape is mathematically modelled by the logistic function:

$$f(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

- The logistic function is the inverse of the logit, i.e. the log-transformed odds

$$f^{-1}(x) = \ln \frac{x}{1 - x}$$

- The logistic function deviates slightly from the cumulative distribution of the normal curve (a.k.a. the inverse probit)



- The s-curve is mathematically defined; it follows a parametric trajectory
- Add parameters α (intercept) and β (slope):

$$y = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

$$\ln \frac{y}{1-y} = \alpha + \beta x$$

$$\text{logit}\ (y) = \alpha + \beta x$$

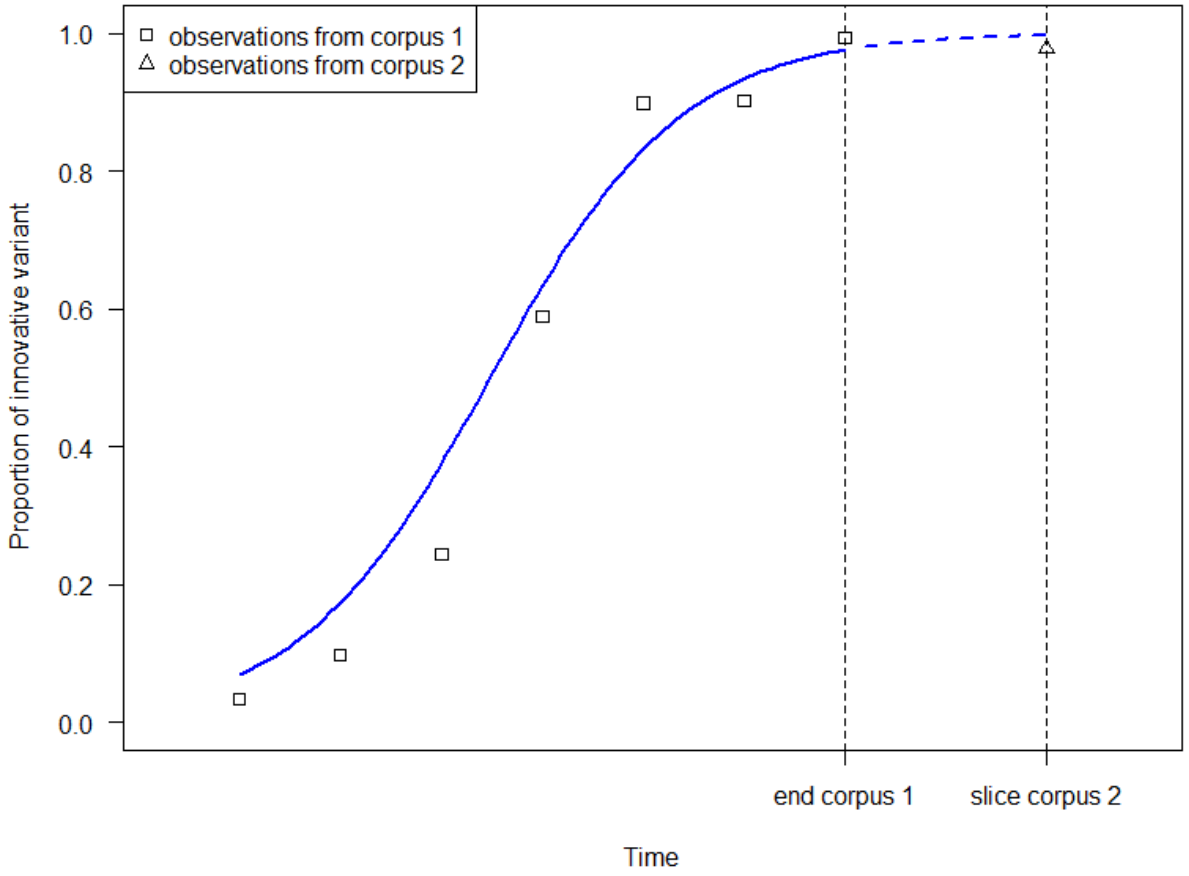
- The latter equation has a link function for the generalised linear model
- y = probability, in function of predictor x = time
- Iterative process of 'maximum likelihood estimation' to find optimal α and β
- Different predictors can be added, with different βs, for multiple regression:

$$\text{logit}\ (y) = \alpha + \sum_{i=0}^{n} \beta_i x_i$$

- If we find α and β, we can plug in any time, and get the probability (or proportion).

$$\text{logit}\ (y) = \alpha + \beta x$$

- We can then look whether this proportion is in line with observed corpus values.



Case study 1: Hortative in Dutch

(1) OBJECT-CX

Laat	*ons*	*naar*	*het*	*strand*	*gaan*
let	us	to	the	beach	go

'let's go to the beach'

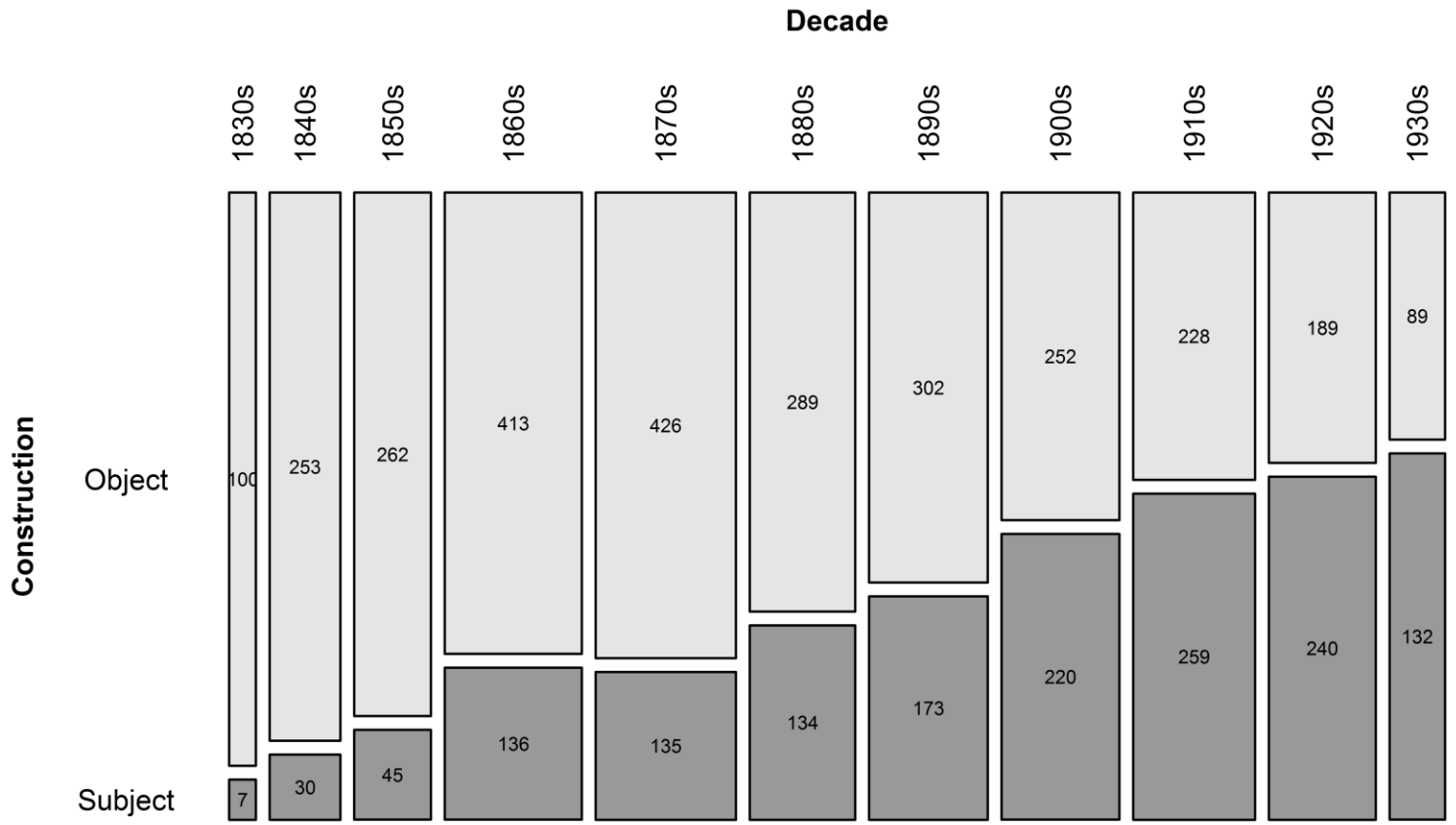
(2) SUBJECT-CX

Laten	*we*	*naar*	*het*	*strand*	*gaan*
let	we	to	the	beach	go

'let's go to the beach'

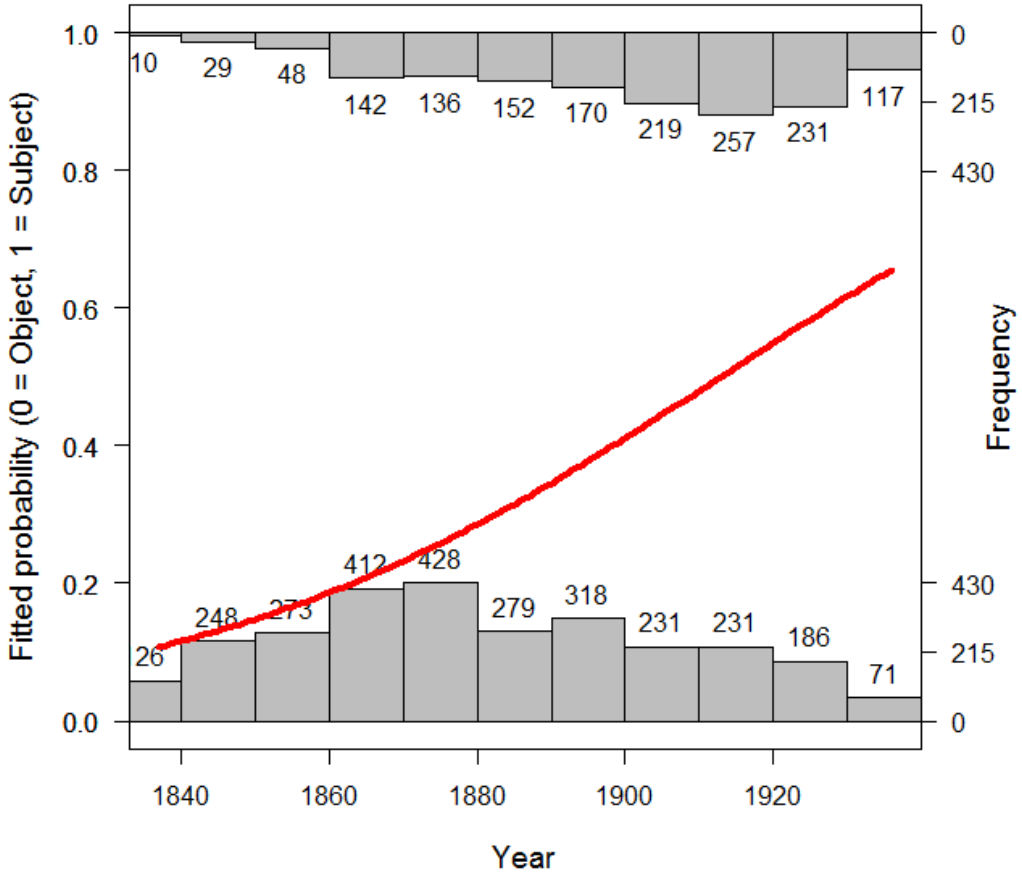
- Source (1): imperative of causative 'allow us to go to the beach' (like in English)
- Source (2): analogy with other mood constructions (Van de Velde 2017)

Historical development: gradual shift from (1) → (2)



→ Gradual rise of innovative subject construction (light grey) borne out by corpus study (*De Gids,* n = 4314)

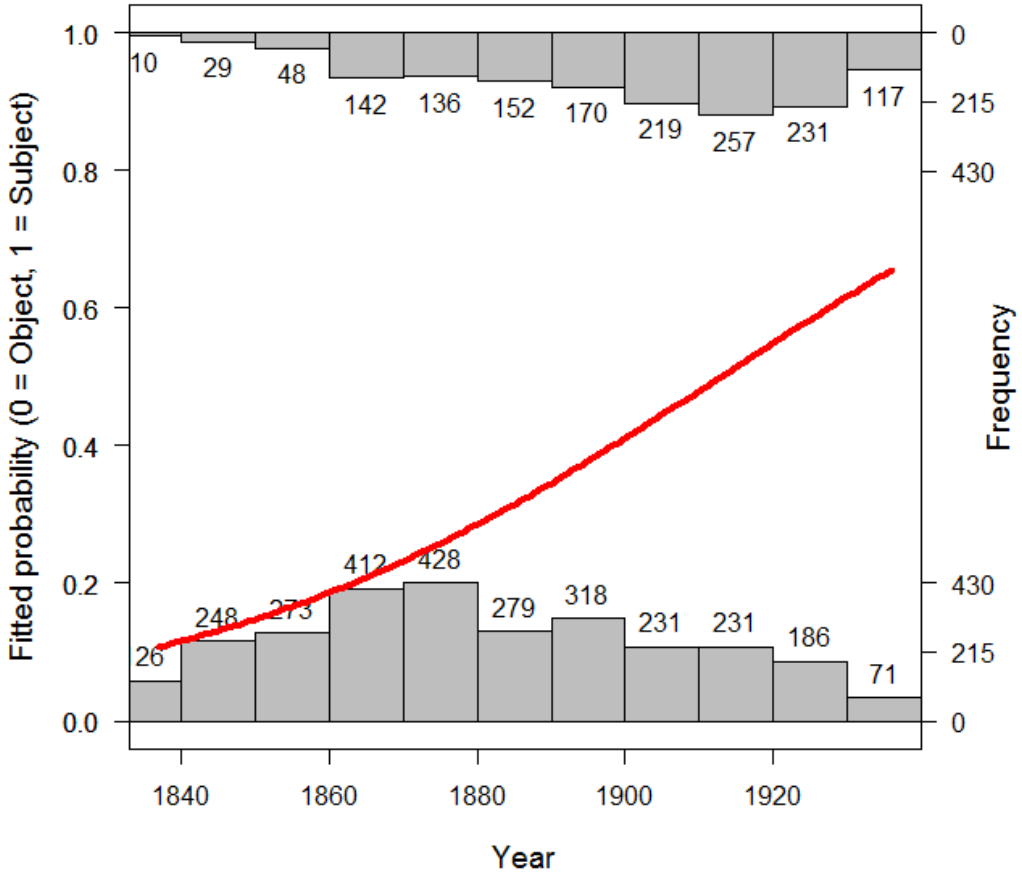
Add trendline (logistic regression)



$$y = \frac{1}{1 + e^{-(\alpha+\beta x)}}$$

$$\alpha = -53.239865$$
$$\beta = 0.027831$$

Add trendline (logistic regression)

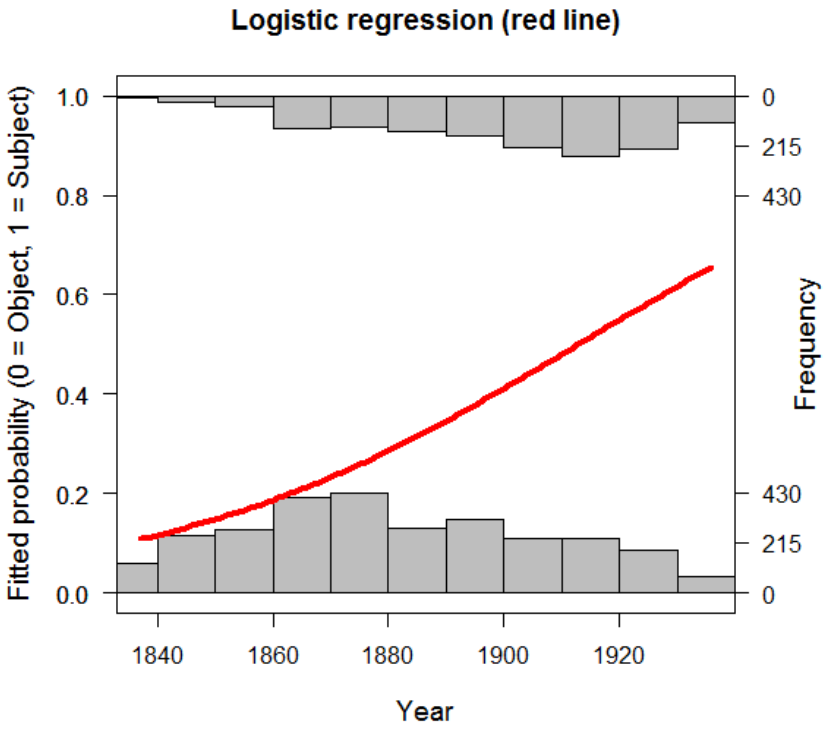


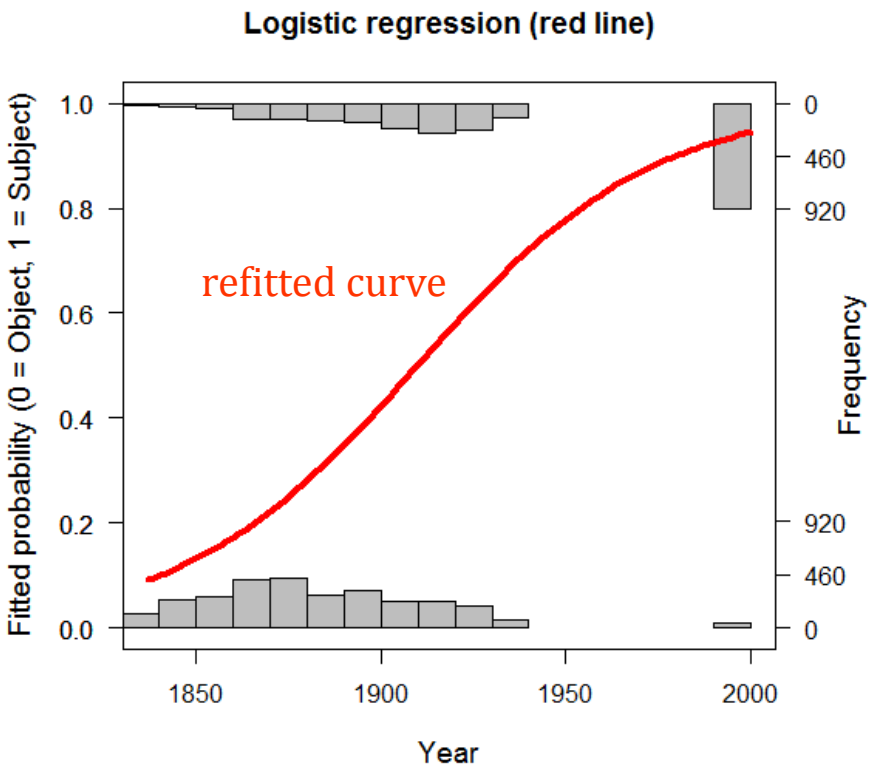
$$y = \frac{1}{1 + e^{-(\alpha+\beta x)}}$$

$\alpha = -53.239865$
$\beta = 0.027831$

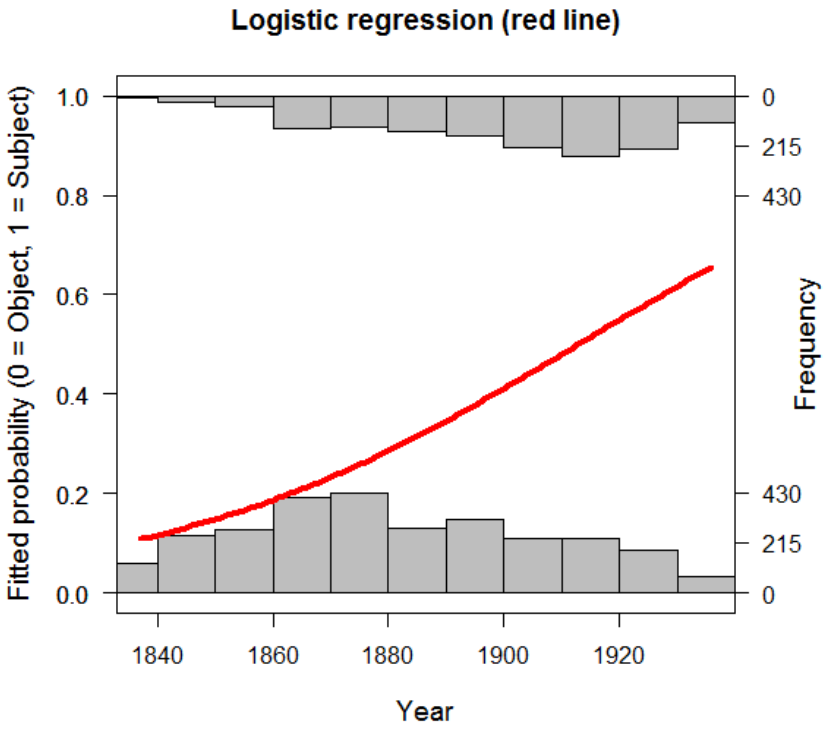
Fitted probability for subject construction in the year 2000 (out of the bounds of the corpus) ≈ 0.92

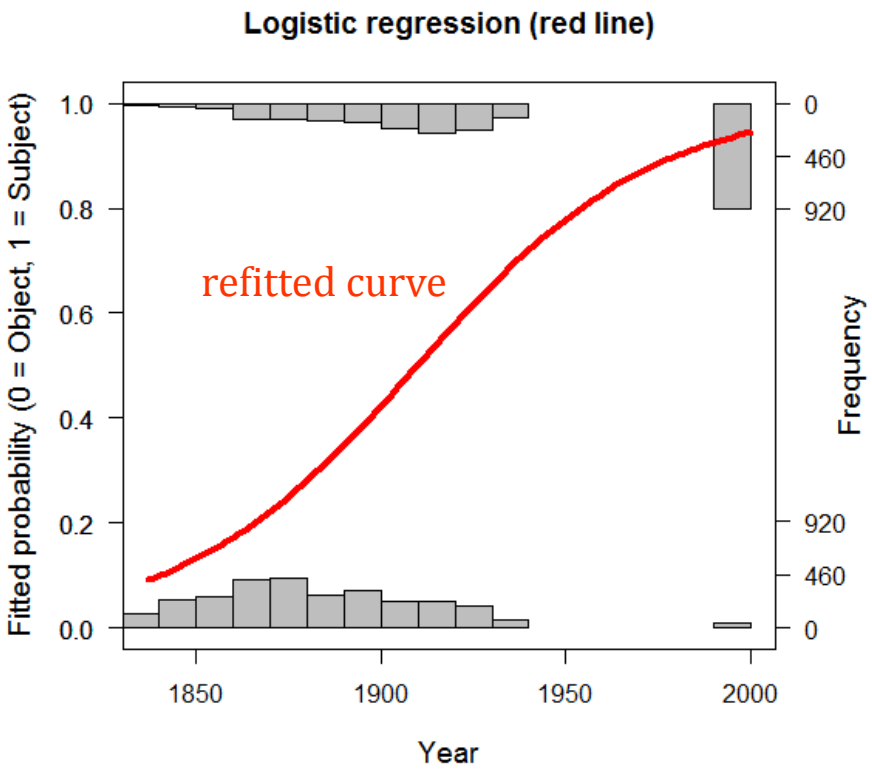
How good is the fit?
→ turn to another corpus, situated in the future



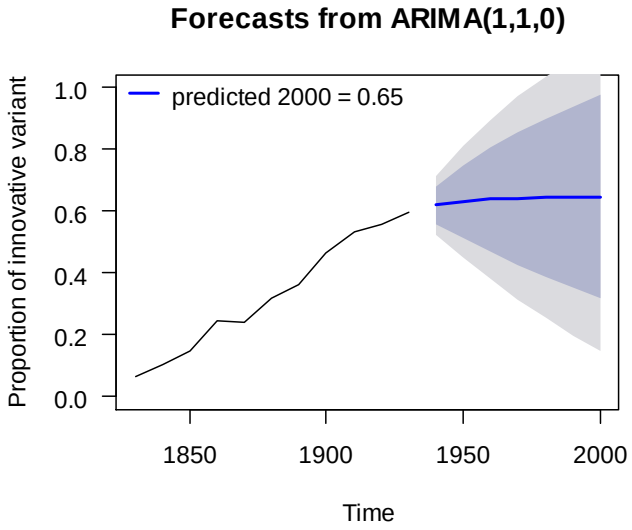


Predicted: 0.92
Found: 0.96
→ only 0.04 off
in *another* corpus (Twente News Corpus),
with a 60+ year gap





Predicted (by old curve): 0.92
Found: 0.96
→ only 0.04 off
in *another* corpus (Twente News Corpus),
with a 60+ year gap



Case study 2: Predeterminer in Dutch

(1) PREDETERMINER-CX

al	*de*	*mensen*
all	the	people

(2) DETERMINER-CX

alle	*mensen*
all	people

Both constructions have syntactic cognates in English

Historical development: gradual shift from (1) → (2) (Van de Velde 2014)

Case study 2: Predeterminer in Dutch

(1) PREDETERMINER-CX

al	*de*	*mensen*
all	the	people

(2) DETERMINER-CX

alle	*mensen*
all	people

Both constructions have syntactic cognates in English

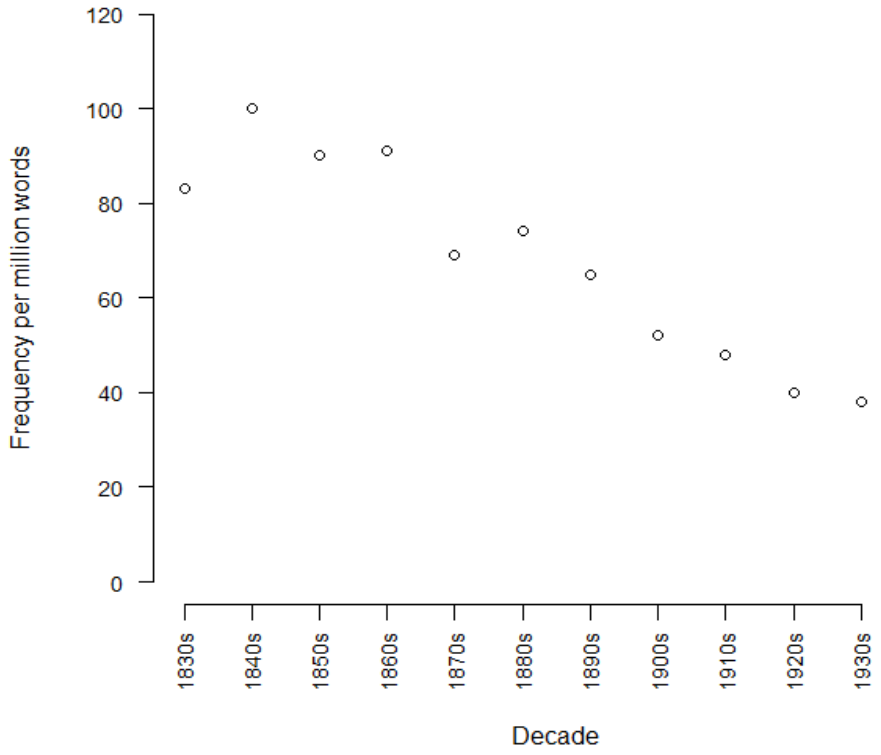
Historical development: gradual shift from (1) → (2) (Van de Velde 2014)

In principle, we could follow the same procedure, and treat (1) vs. (2) as a classical alternation. However, there are other mutants:

(3) FLOATING CX

De	*mensen*	*hebben*	*alle(maal)*	*gewerkt*
the	people	have	all	worked

Another procedure is to just look at the frequency per million words of the old variant

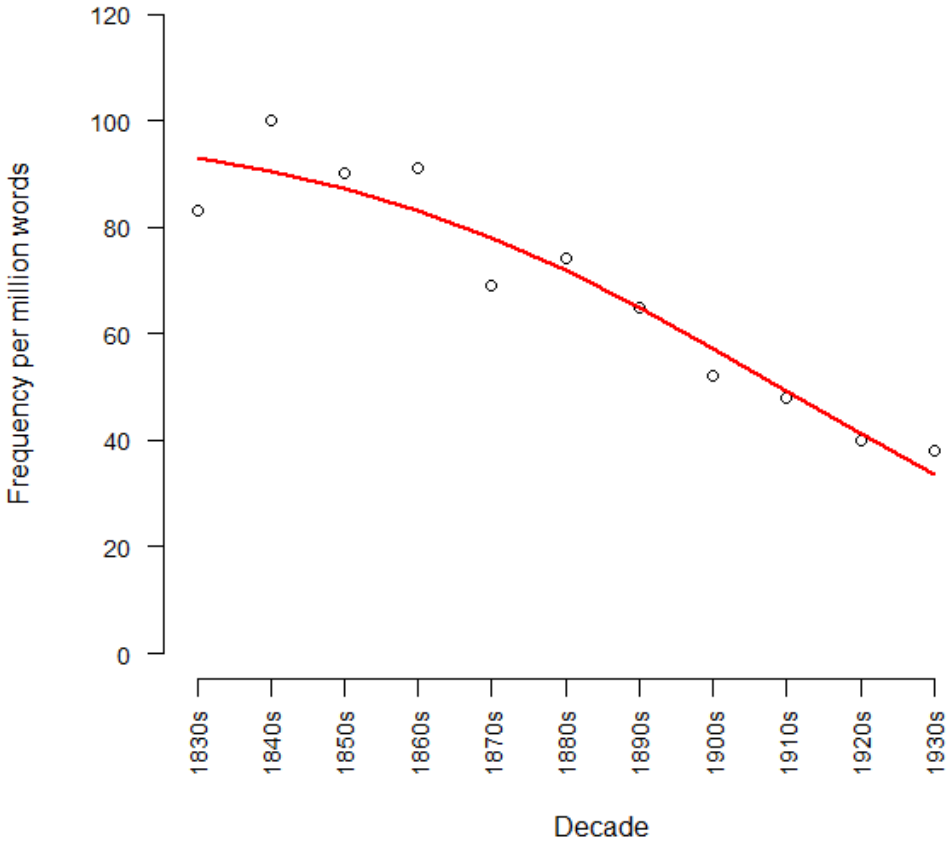


Here we have a construction on its way out. Does that follow an S-curve as well?

One could be tempted to fit a linear regression line through these datapoints, but that does not do justice to the hypothesised S-curve

We can also fit an S-curve through the observed points, treating the (rounded) highest attested frequency per million words (100, for the 1840s) as the number of successes and the difference between that number and the observed (rounded) frequency in each decade as the number of failures.

The effect of de decade is highly significant ($p < 0.0001$)



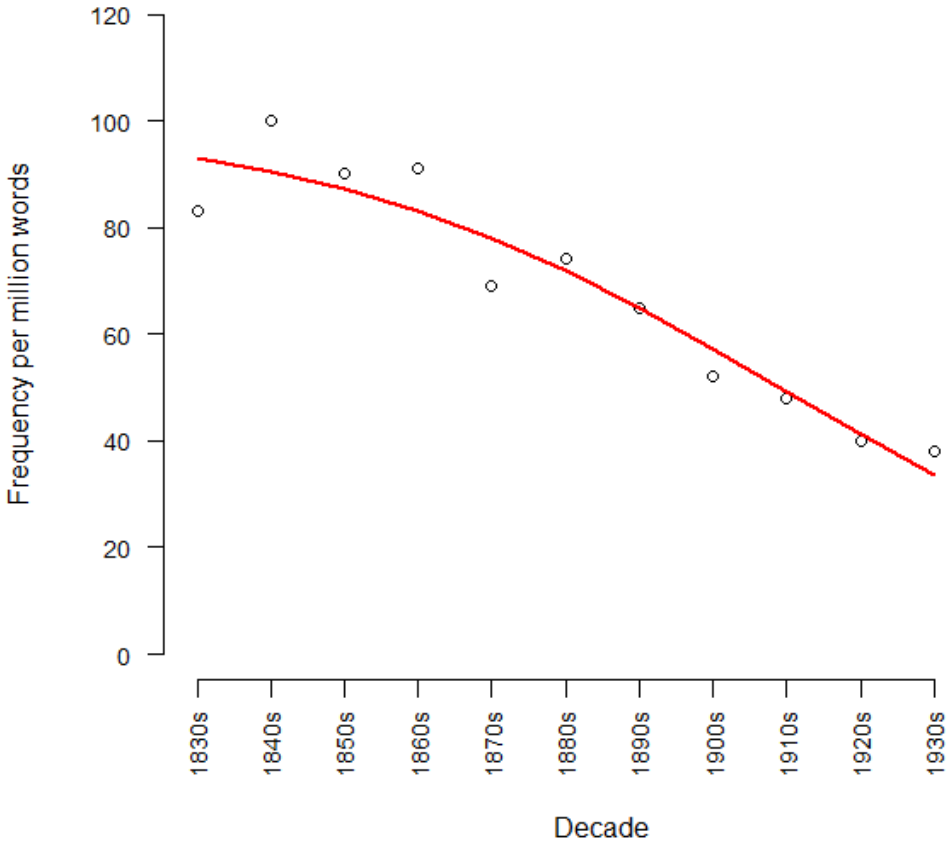
What will happen if we project that line onto a later data, in another corpus? Will the procedure still work? Given that:

(a) we have a waning, instead of growing variant

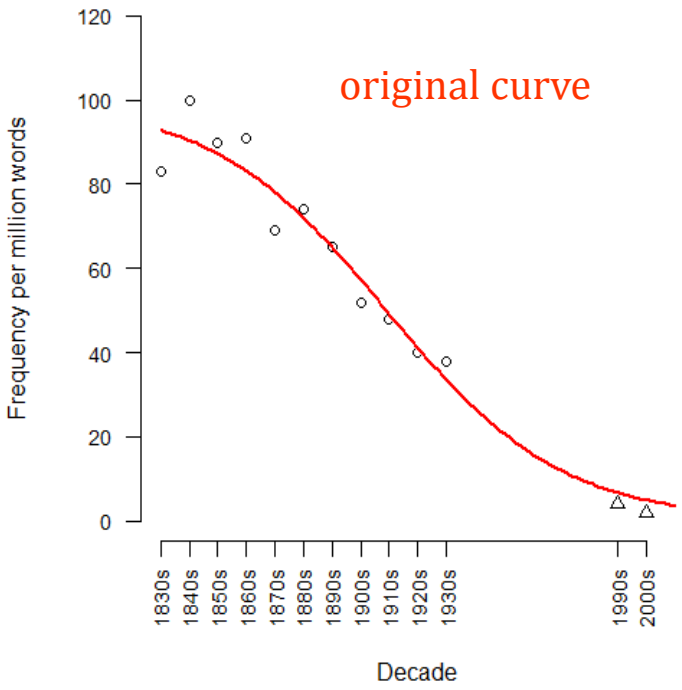
(b) we have much fewer datapoints to fit the curve (n = 11)

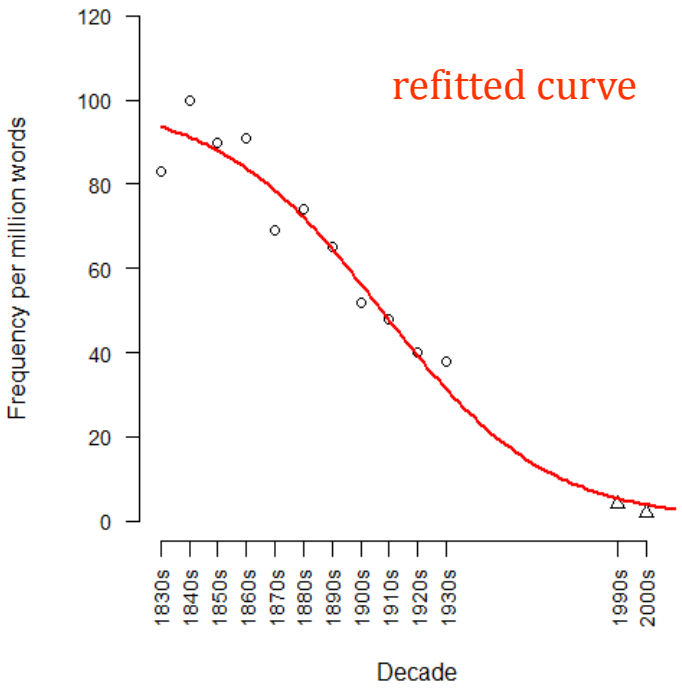
(c) we have another corpus (Literom)

(d) we are going to fit it through two datapoints in the future (1990s and 2000s)



It works:





Curve fitted on 1830s to 1930s only

Curve fitted on 1830s to 1930s AND 1990s and 2000s

→ almost exactly the same

Case study 3: Complex prepositions in Dutch

(1) $[_{PP}$ *in*$_P$ $[_{NP}$*het*$_D$ *kader*$_N$ $[_{PP}$ *van*$_P$ $[_{NP}$ …]]]]
in the frame of

(2) $[_{PP}$ *in-het-kader-van*$_P$ $[_{NP}$ …]]
in-the-frame-of

Complex prepositions on the basis of a [P (D) N P] pattern. Same in English: *in view of*, *in spite of*, *with regard to*, *by dint of* ... (Hoffmann 2005)

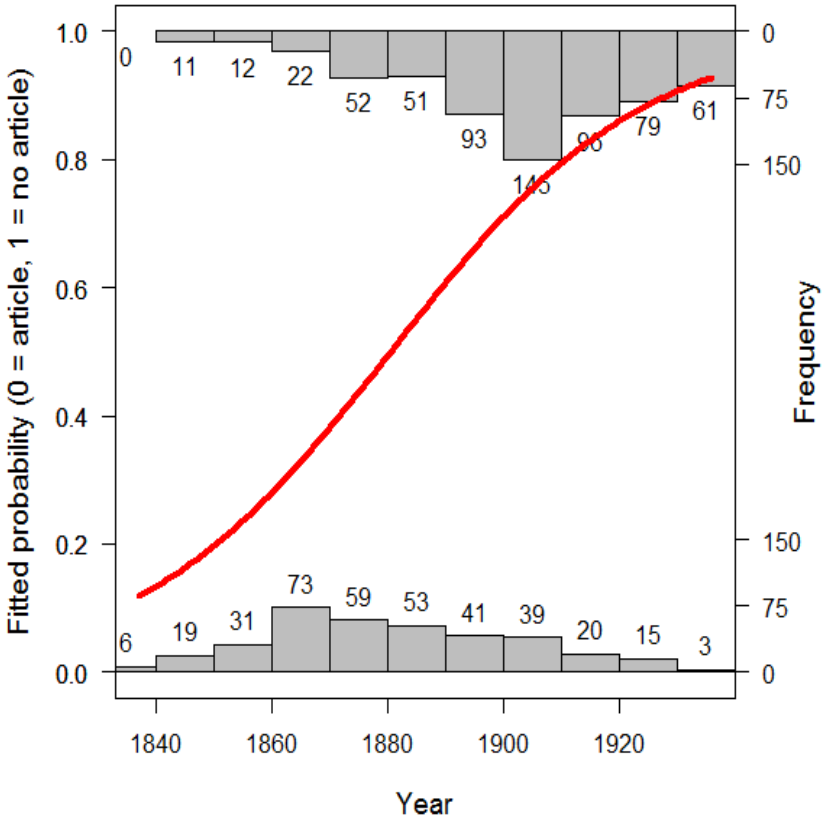
Examples abound: *aan de hand van*, *met het oog op*, *met betrekking tot* ... (Loonen 2003)

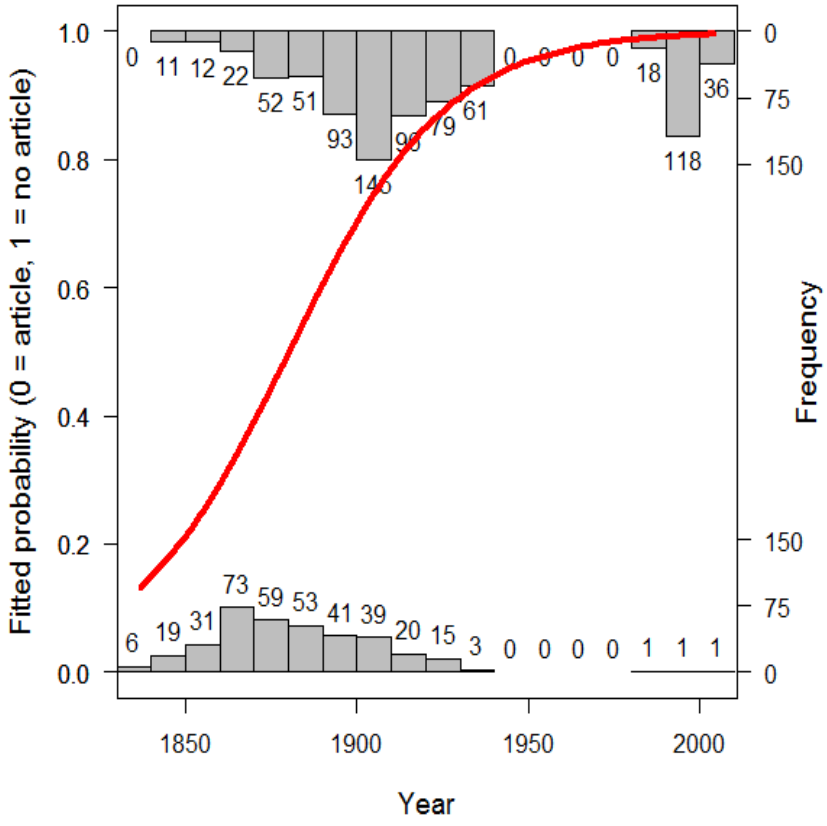
These complex prepositions have the tendency to drop the article, in a process of 'decategorialisation' (Van der Horst 2004; Vranjes 2012; Hüning 2014)

(1) *onder de leiding van* ('under the command of')

(2) *onder Ø leiding van* ('under command of')

Historical development: gradual shift from (1) → (2) (Vranjes 2012)





$$y = \frac{1}{1 + e^{-(\alpha+\beta x)}}$$

$\alpha = -86.606510$
$\beta = 0.046054$

	Predicted	Observed
1999	0.99	0.98
2000	1.00	0.98

Conclusion

- The trajectory of change is remarkably predictable
  - in different changes
  - with different validation corpora (TwNC, Literom)
  - with phenomena on the rise and phenomena on decline
  - with alternations and with normalised frequency measures
  - with fewer and with more datapoints

- I have only looked at Late Modern Dutch changes
  - method is less likely to work for long-time changes (successive S-curves..., see Nevalainen 2015)
  - Method is less likely to work in times of sudden demographic upheaval (see also De Smet, Beuls, Pijpops & Van de Velde, this conference)

Bauer, L. 1994. *Watching English change: an introduction to the study of linguistic change in standard Englishes in the twentieth century*. London: Longman.

Blythe, R.A. & W. Croft. 2012. 'S-curves and the mechanisms of propagation in language change'. *Language* 88(2): 269-304.

Croft, W. 2000. *Explaining language change: an evolutionary approach*. Harlow: Longman.

Denison, D. 2003. 'Log(ist)ic and simplistic S-curves'. In: R. Hickey (ed.), *Motives for language change*. Cambridge: Cambridge University Press. 54-70.

De Saussure, F. 1955 [1916]. *Cours de linguistique générale. Publié par Charles Bally et Albert Sechehaye avec la collaboration de Albert Riedlinger*. Paris: Payot.

Hoffmann, S. 2004. 'Are low-frequency complex prepositions grammaticalized? On the limits of corpus data - and the importance of intuition'. In: H. Lindquist & C. Mair (ed.) *Corpus approaches to grammaticalization in English*. Amsterdam: John Benjamins. 171-210.

Hüning, M. 2014. 'Over complexe preposities en convergentie'. In: F. Van de Velde, H. Smessaert, F. Van Eynde & S. Verbrugge (eds.), *Patroon en argument. Een dubbelfeestbundel bij het emeritaat van William Van Belle en Joop van der Horst*. Leuven: Leuven University Press. 433-445.

Kroch, A.S. 1989. 'Reflexes of grammar in patterns of language change'. *Language Variation and Change* 1: 199-244.

Loonen, Nard. 2003. *Stante pede gaande van dichtbij langs AF bestemming.* PhD Diss. Utrecht University.

Nevalainen, T. 2015. 'Descriptive adequacy of the S-curve model in diachronic studies of language change'. In: C. Sanchez-Stockhammer (ed.), *Can we predict linguistic change?* Special issue of *Varieng: Studies in Variation, Contacts and Change in English*. Volume 16.

Pintzuk, S. 2003. 'Variationist approaches to syntactic change'. In: B.D. Joseph & R.D. Janda (eds.), *The handbook of historical linguistics*. Oxford: Blackwell. 509-528.

Sanchez-Stockhammer, C. 2015. 'Can we predict linguistic change? An introduction'. In: C. Sanchez-Stockhammer (ed.), *Can we predict linguistic change?* Special issue *of Varieng: Studies in Variation, Contacts and Change in English*. Volume 16.

Van der Horst, J. 2004. 'Met (het) oog op morgen: de verdwijning van lidwoorden uit vaste verbindingen'. *Onze Taal* 73: 96-98.

Van de Velde, F. 2014. 'Nederlandse predeterminatoren als levend fossiel'. *Nederlandse Taalkunde* 19(1): 87-103.

Van de Velde, F. 2017. 'Understanding grammar at the community level requires a diachronic perspective. Evidence from four case studies'. *Nederlandse Taalkunde* 22(1): 47-74.

Vranjes, Jelena. 2012. *Voorzetseluitdrukkingen in het Nederlands sinds de 16de eeuw: een diachroon corpusonderzoek*. Ma-thesis, University of Leuven.

Weinreich, U., W. Labov & M. Herzog. 1968. 'Empirical foundations for a theory of language change'. In: W.P. Lehmann & Y. Malkiel (eds.), *Directions for historical linguistics*. Austin: University of Texas Press. 95-188.